Final Assignment Report

Name : Maad Saifuddin

Online Shopper Intention Data Set

Introduction : The data set consists of feature vectors belonging to 12,330 sessions. The data set was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period.

Objective : Finding customer buying behaviour.

Features of our Data set are :

Administrative: Administrative Value numeric

Administrative_Duration: Duration in Administrative Page numeric

Informational: Informational Value numeric

Informational_Duration: Duration in Informational Page numeric

ProductRelated: Product Related Value numeric

ProductRelated_Duration: Duration in Product Related Page numeric

BounceRates: Bounce Rates of a web page numeric

ExitRates: Exit rate of a web page numeric

PageValues: Page values of each web page numeric

SpecialDay: Special days like valentine etc

Month: Month of the year (categorical) eg (jan,feb,mar etc)

OperatingSystems: Operating system used **numeric** from 1 to 8

Browser: Browser used categorical from 1 to 13

Region: Region of the user categorical from 1 to 9

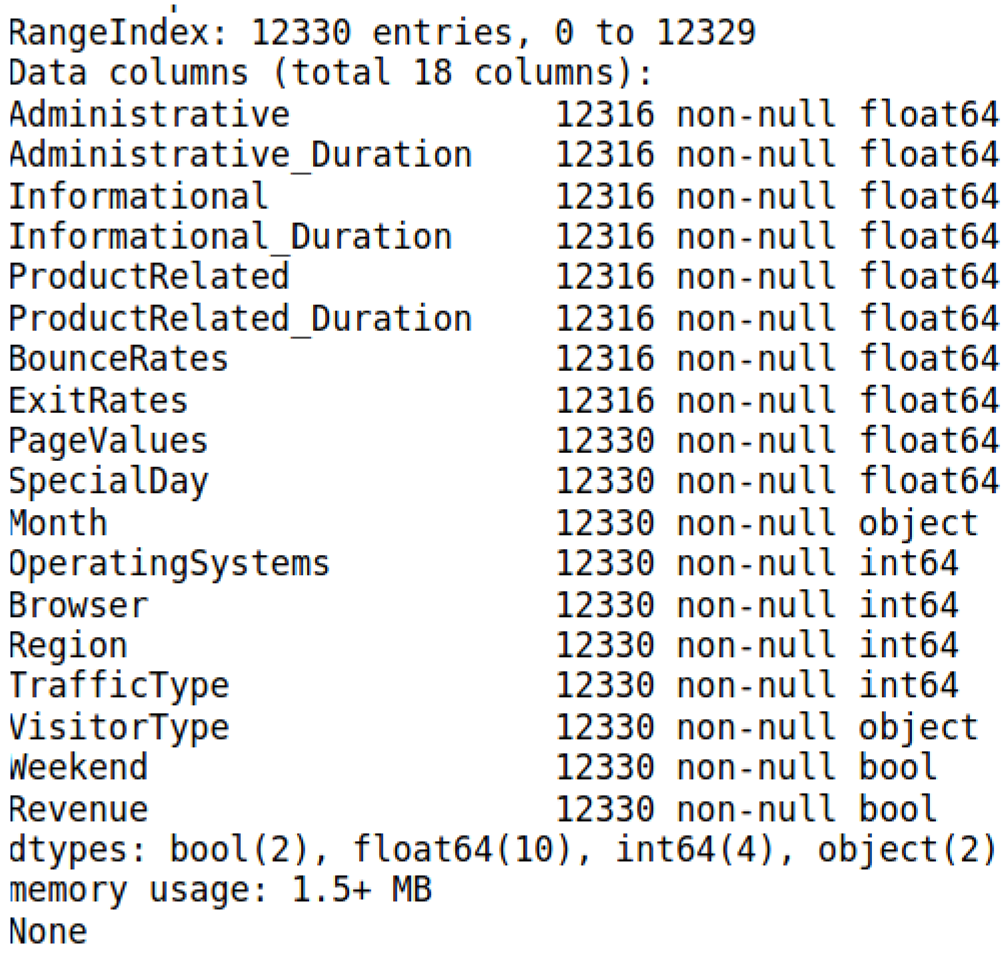
TrafficType: Traffic Type categorical from 1 to 20

VisitorType: Types of Visitor categorical 'Returning_Visitor' 'New_Visitor' 'Other'

Weekend: Weekend or not boolean/categorical true/false

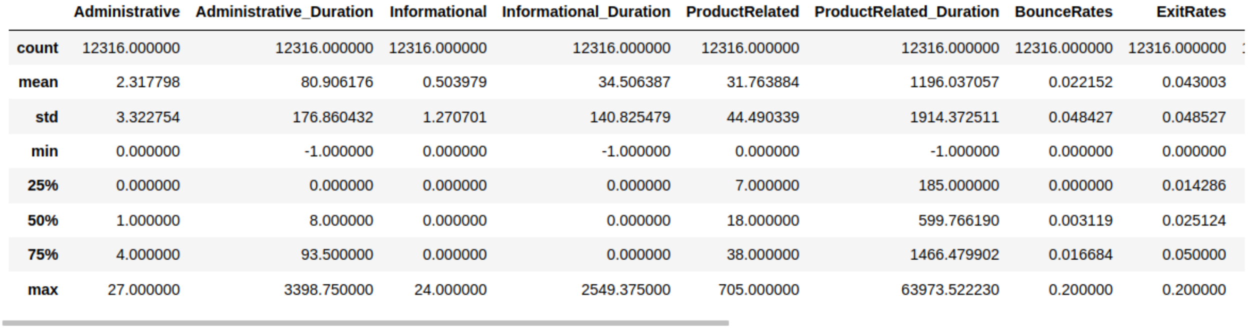
Revenue: Revenue will be generated or not boolean/categorical true/false

Data Wrangling and Preprocessing: The Data set contain 12330 observations and 18 columns the data information are.

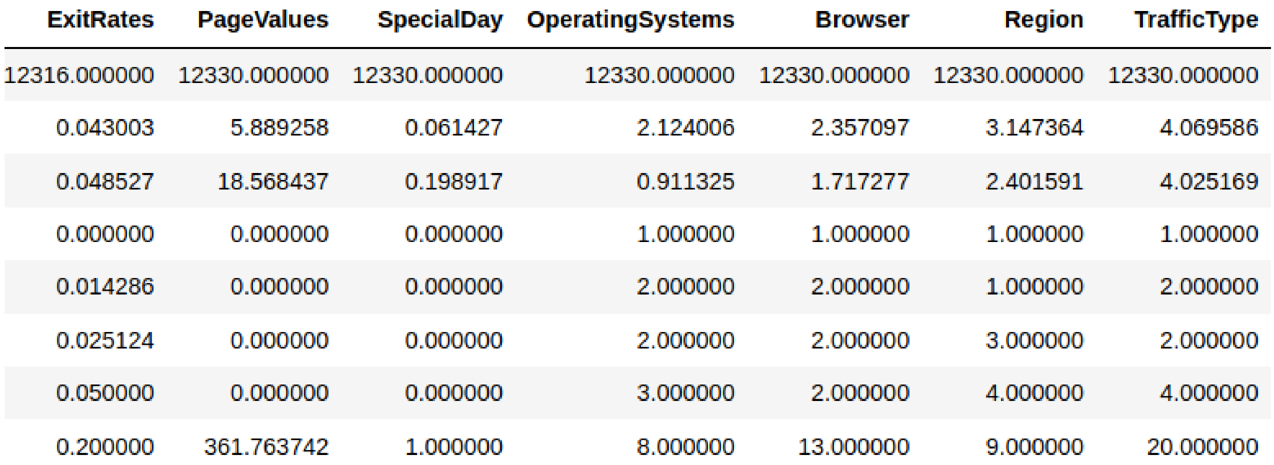


```
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 18 columns):
Administrative              12316 non-null float64
Administrative_Duration     12316 non-null float64
Informational               12316 non-null float64
Informational_Duration      12316 non-null float64
ProductRelated              12316 non-null float64
ProductRelated_Duration     12316 non-null float64
BounceRates                 12316 non-null float64
ExitRates                   12316 non-null float64
PageValues                  12330 non-null float64
SpecialDay                  12330 non-null float64
Month                       12330 non-null object
OperatingSystems            12330 non-null int64
Browser                     12330 non-null int64
Region                      12330 non-null int64
TrafficType                 12330 non-null int64
VisitorType                 12330 non-null object
Weekend                     12330 non-null bool
Revenue                     12330 non-null bool
dtypes: bool(2), float64(10), int64(4), object(2)
memory usage: 1.5+ MB
None
```

The Statistics of online dataset is :

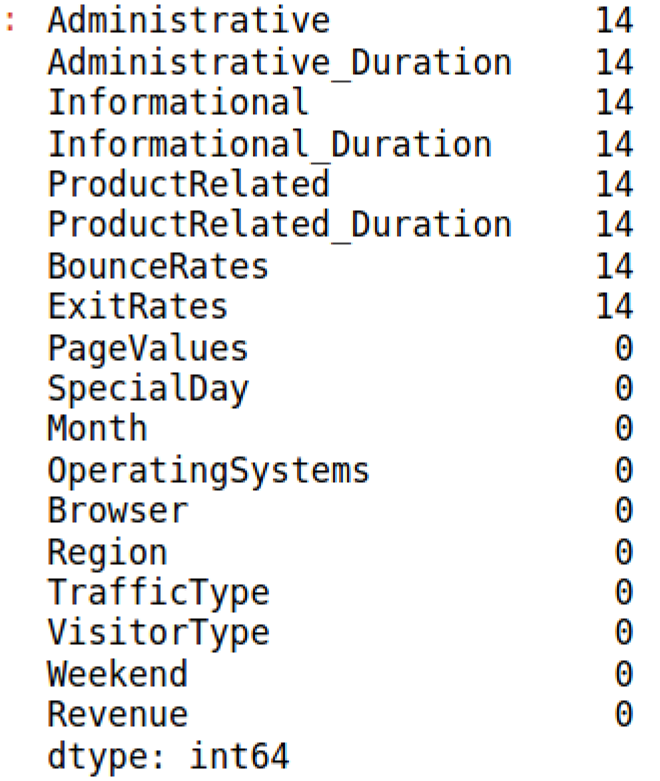


	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates
count	12316.000000	12316.000000	12316.000000	12316.000000	12316.000000	12316.000000	12316.000000	12316.000000
mean	2.317798	80.906176	0.503979	34.506387	31.763884	1196.037057	0.022152	0.043003
std	3.322754	176.860432	1.270701	140.825479	44.490339	1914.372511	0.048427	0.048527
min	0.000000	-1.000000	0.000000	-1.000000	0.000000	-1.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	7.000000	185.000000	0.000000	0.014286
50%	1.000000	8.000000	0.000000	0.000000	18.000000	599.766190	0.003119	0.025124
75%	4.000000	93.500000	0.000000	0.000000	38.000000	1466.479902	0.016684	0.050000
max	27.000000	3398.750000	24.000000	2549.375000	705.000000	63973.522230	0.200000	0.200000



ExitRates	PageValues	SpecialDay	OperatingSystems	Browser	Region	TrafficType
12316.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000
0.043003	5.889258	0.061427	2.124006	2.357097	3.147364	4.069586
0.048527	18.568437	0.198917	0.911325	1.717277	2.401591	4.025169
0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
0.014286	0.000000	0.000000	2.000000	2.000000	1.000000	2.000000
0.025124	0.000000	0.000000	2.000000	2.000000	3.000000	2.000000
0.050000	0.000000	0.000000	3.000000	2.000000	4.000000	4.000000
0.200000	361.763742	1.000000	8.000000	13.000000	9.000000	20.000000

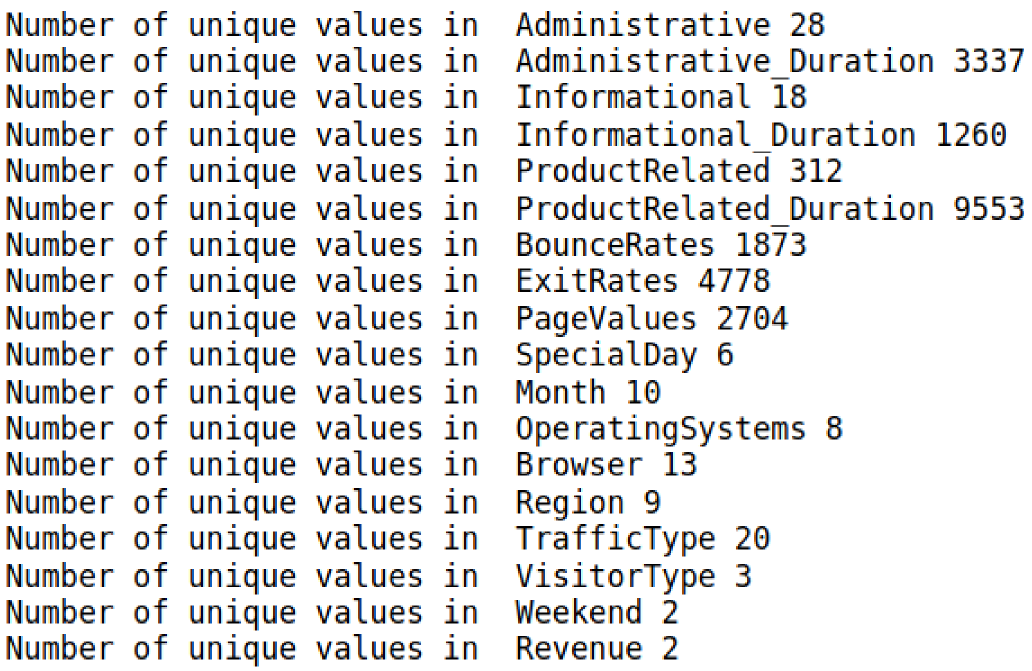
The four features of the data set contain null values .



```
: Administrative             14
  Administrative_Duration    14
  Informational              14
  Informational_Duration     14
  ProductRelated             14
  ProductRelated_Duration    14
  BounceRates                14
  ExitRates                  14
  PageValues                  0
  SpecialDay                  0
  Month                       0
  OperatingSystems            0
  Browser                     0
  Region                      0
  TrafficType                 0
  VisitorType                 0
  Weekend                     0
  Revenue                     0
  dtype: int64
```

This is the missing value at random so we remove all the rows from data set which contain null values.
We conclude that all missing values are Not at random because when the Revenue Value is False our missing value is in greater % 0.13 where as when value of Revenue is True Missing Value is 0%.

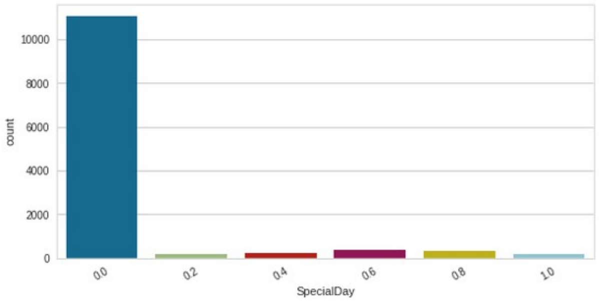
Check for Cardinality:

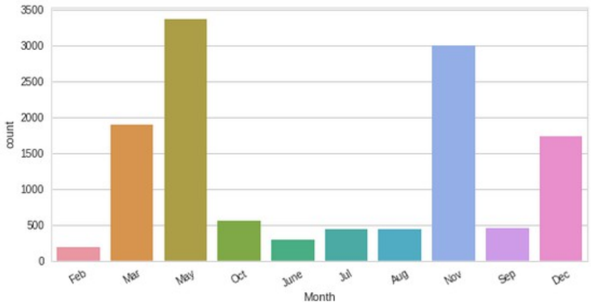


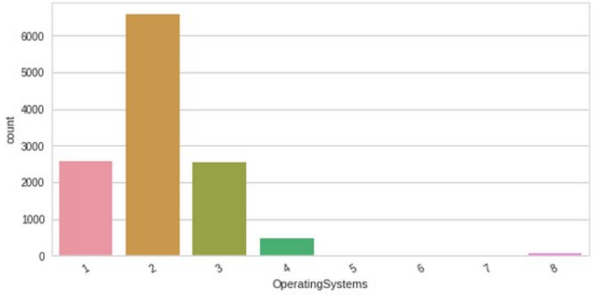
```
Number of unique values in  Administrative 28
Number of unique values in  Administrative_Duration 3337
Number of unique values in  Informational 18
Number of unique values in  Informational_Duration 1260
Number of unique values in  ProductRelated 312
Number of unique values in  ProductRelated_Duration 9553
Number of unique values in  BounceRates 1873
Number of unique values in  ExitRates 4778
Number of unique values in  PageValues 2704
Number of unique values in  SpecialDay 6
Number of unique values in  Month 10
Number of unique values in  OperatingSystems 8
Number of unique values in  Browser 13
Number of unique values in  Region 9
Number of unique values in  TrafficType 20
Number of unique values in  VisitorType 3
Number of unique values in  Weekend 2
Number of unique values in  Revenue 2
```

Our data set contain negative values so we remove negative values from durations features and impute 0.

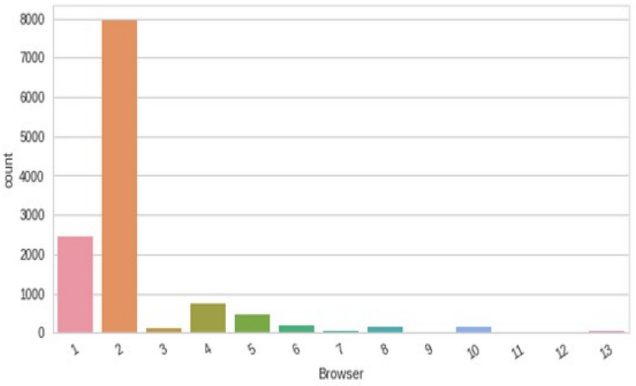
Univariate Analysis :

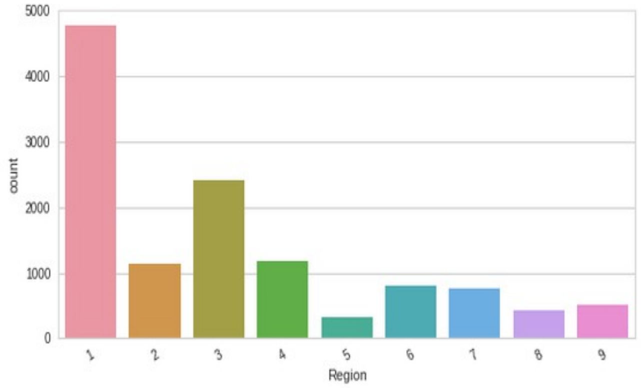


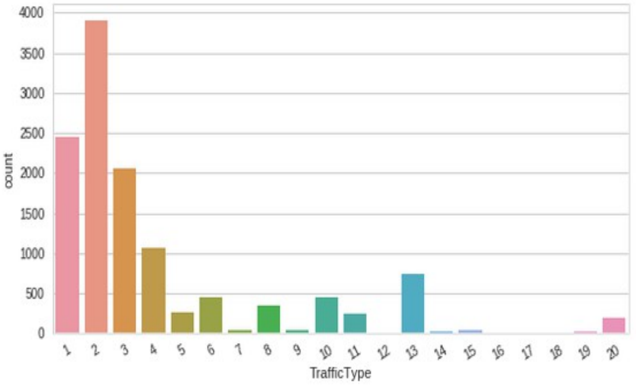


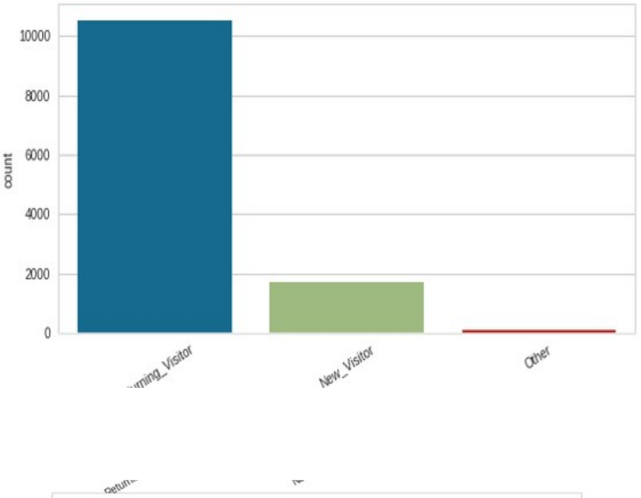


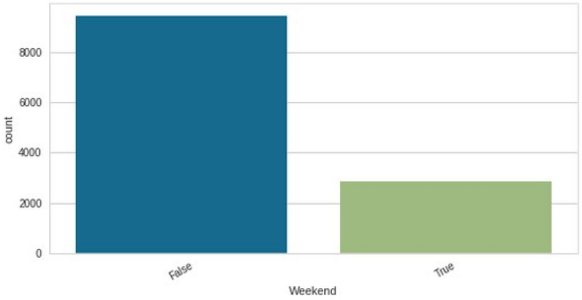
1. Special Day 0.0 occur most time.
2. May and November occur mostly then other months.
3. Number 2 operating system used mostly.

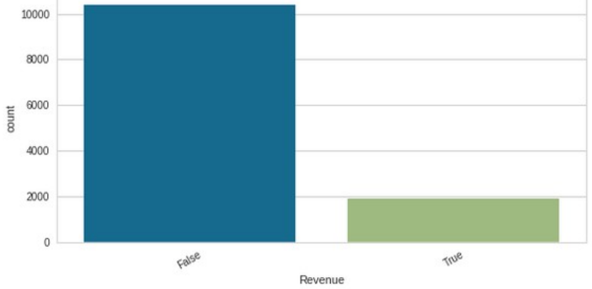








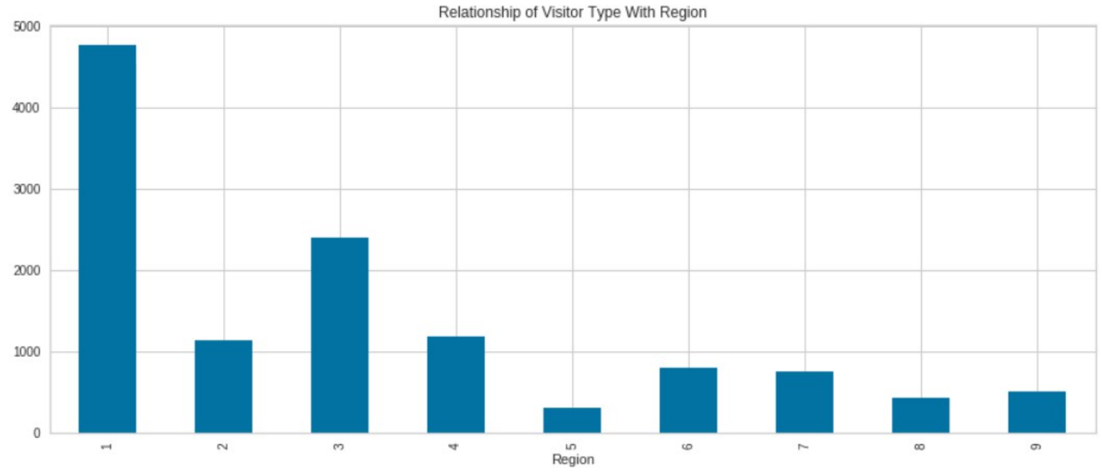




1. Number 2 browser used more than other browser
2. user of region 1 are highest in online search
3. Number 2 traffic type is greater in number than other trafic type
4. Returning visitor are more than other visitor
5. Searches is done more in odd days
6. Mostly the revenue is not generated

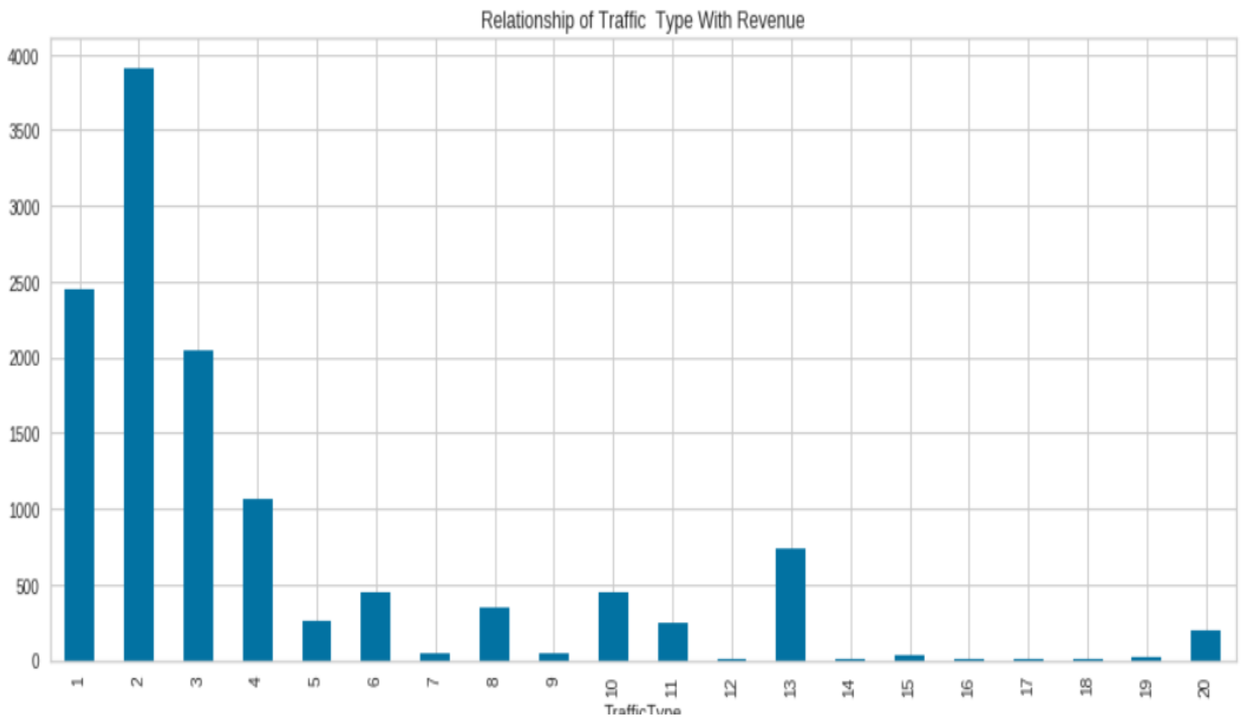
Bivariate Analysis:

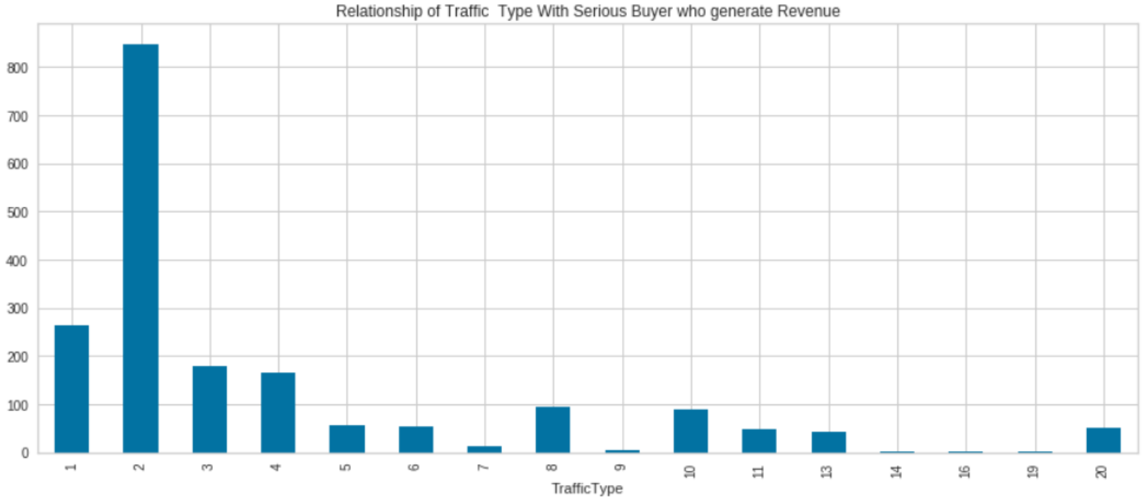
Where does most of the customers visiting the website belonging to? (identify the regions)



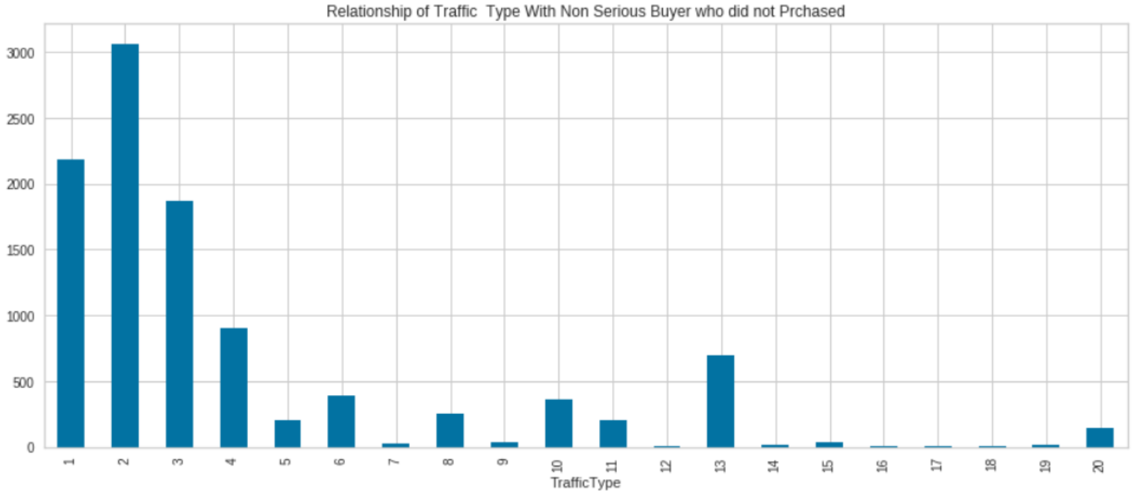
Most of the Customer Visiting Website in from region 1 and 3.

What is the effect of Traffic type on Revenue?





This is graph show that from traffic type 2,1,3,4 mostly the customer visit and purchased.

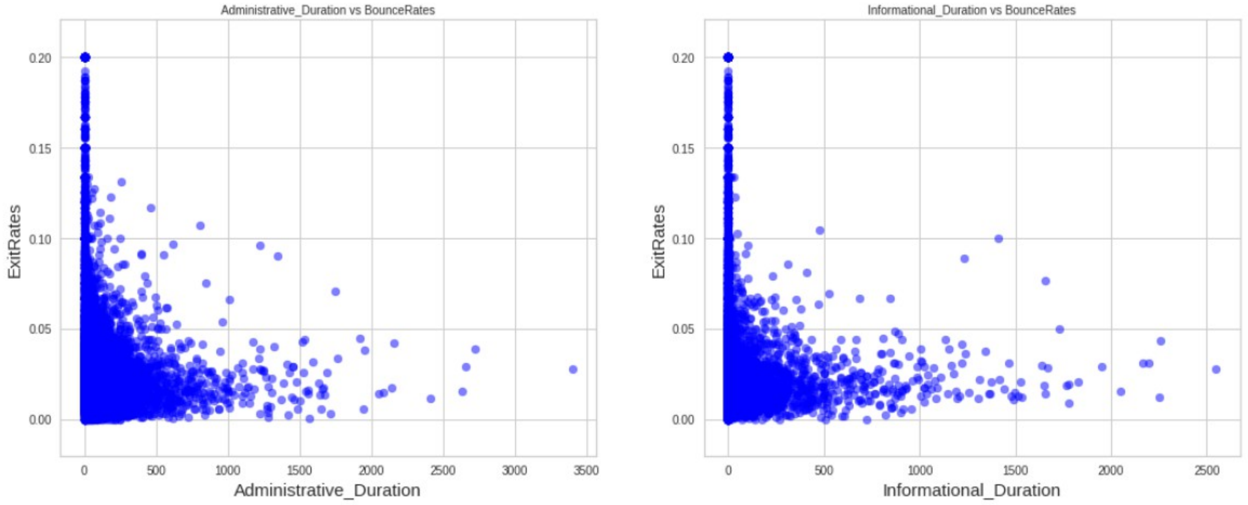


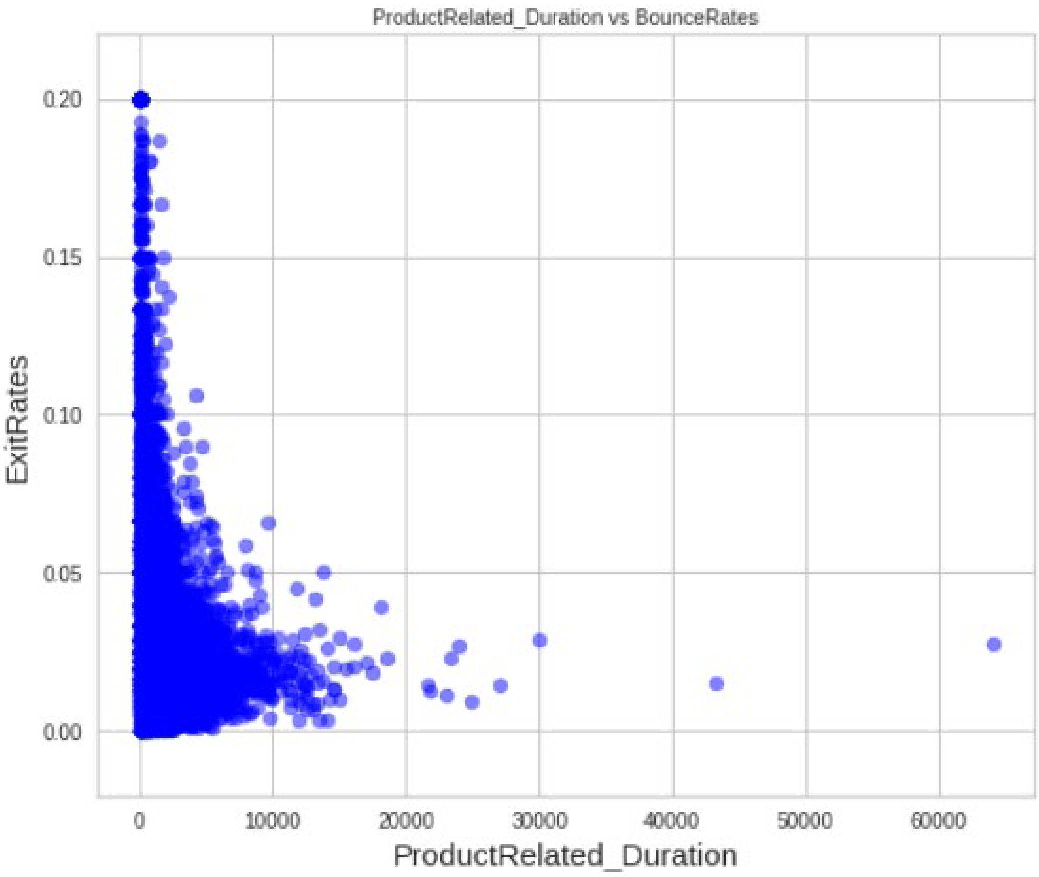
This graph represent that from traffic type 1,2,3,4 customer visit the website but did not purchased.

From both graph we analyze that the purchased is very rare where as buying is not done in greater amount.

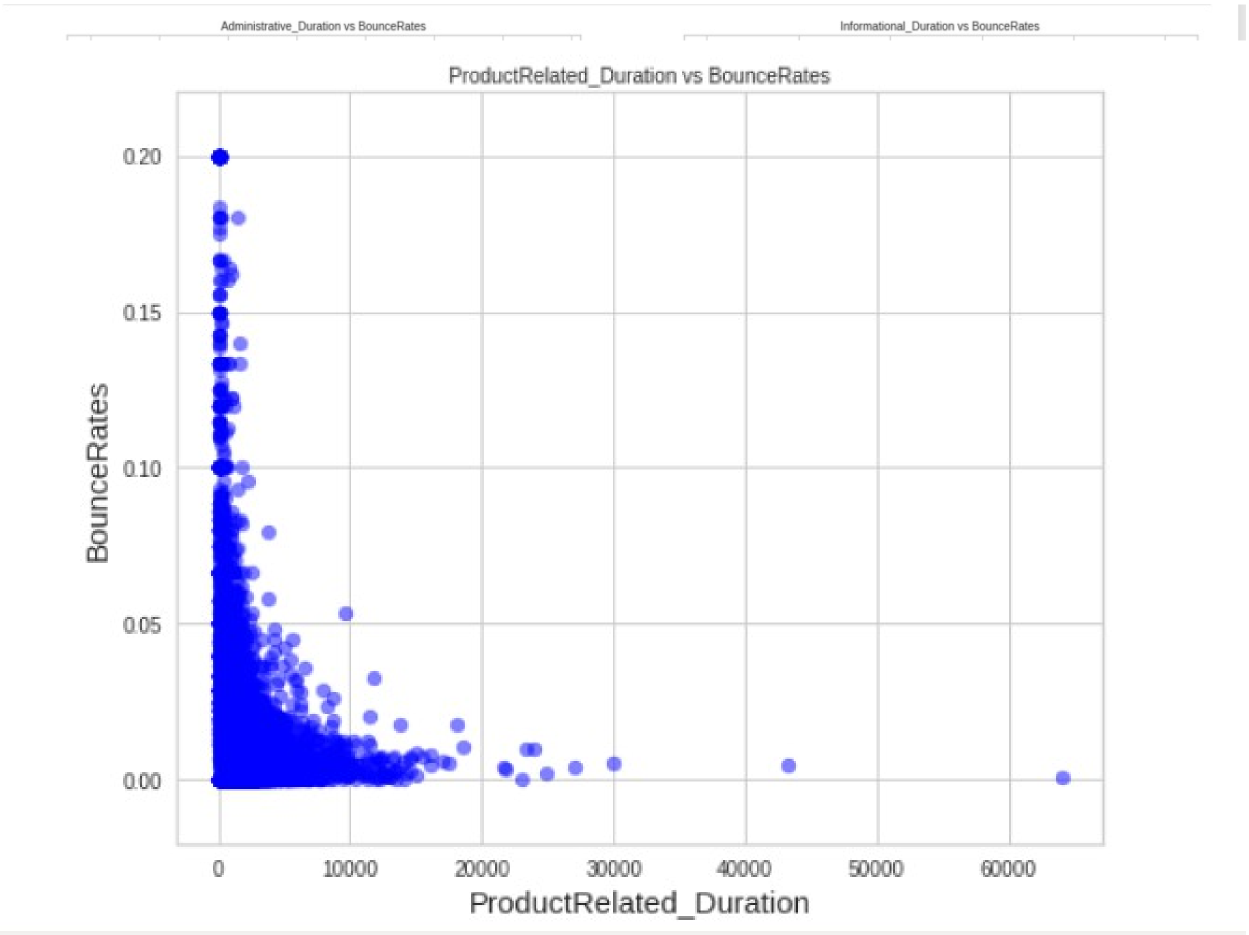
From traffic type 2 Their is balance between purchased done or not.

How does longer duration spent on the website affect the Exit rate?



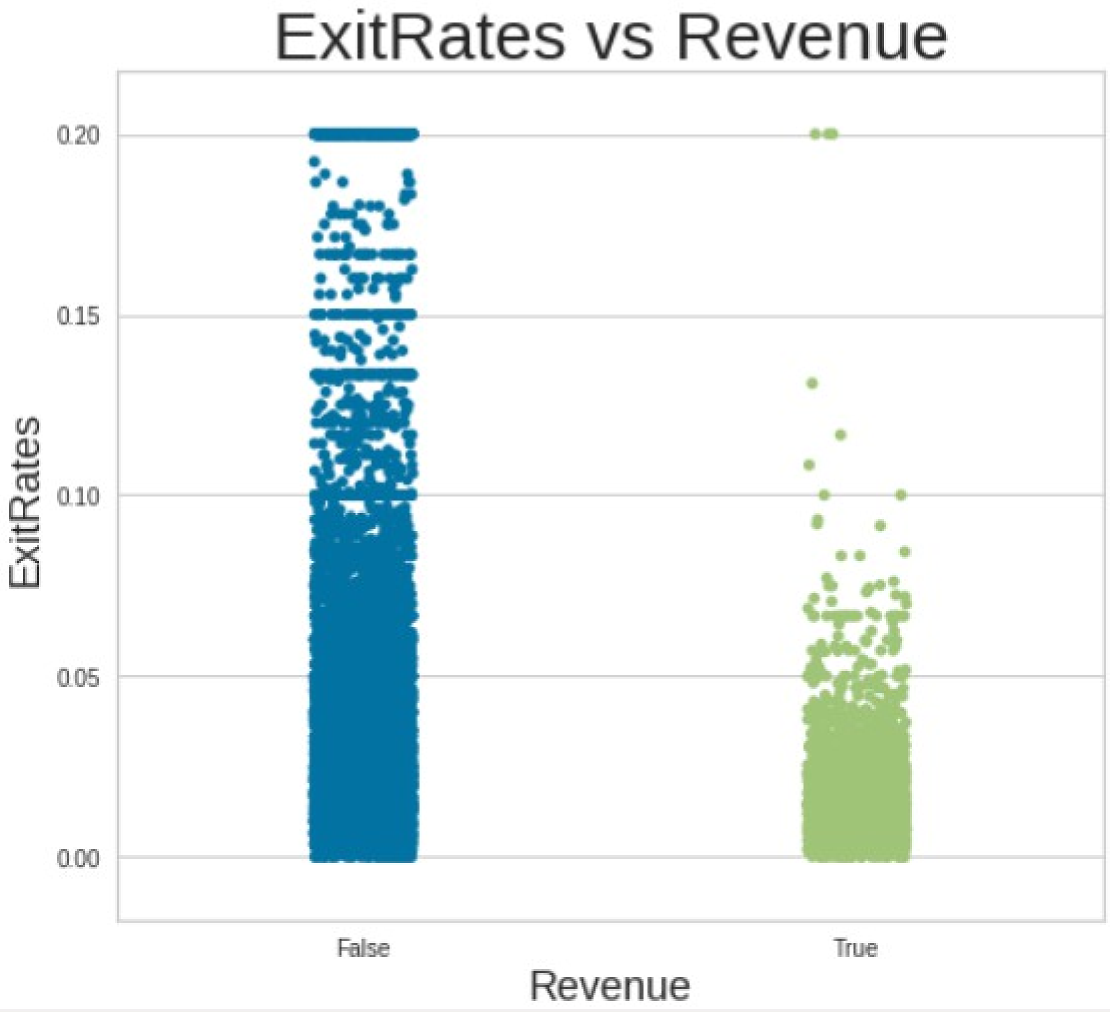


How does longer duration spent on the website affect the bounce rate?

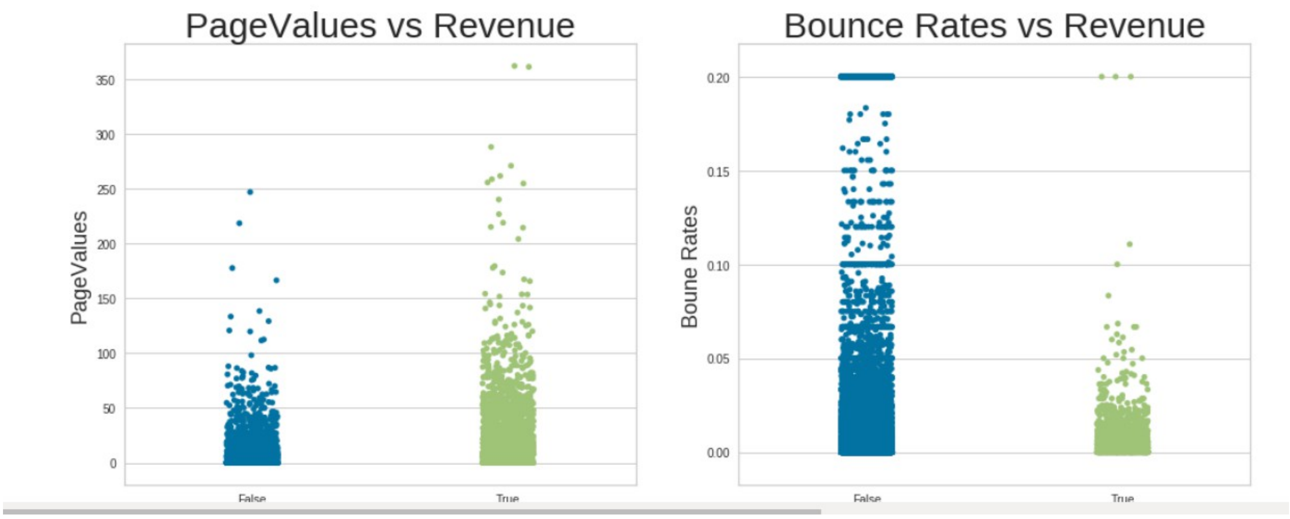


Above graphs represent as when duration increase bounce rate is start reducing.

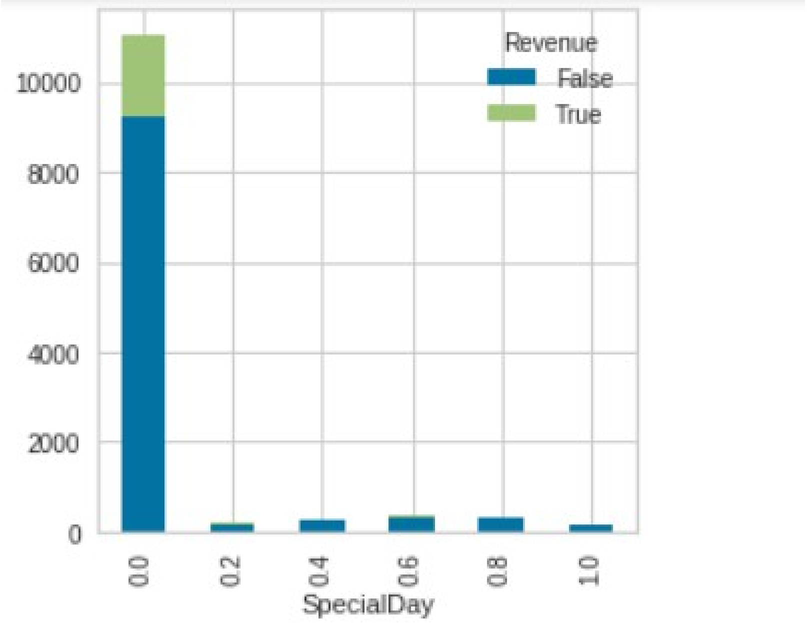
How does exit rate affect revenue?



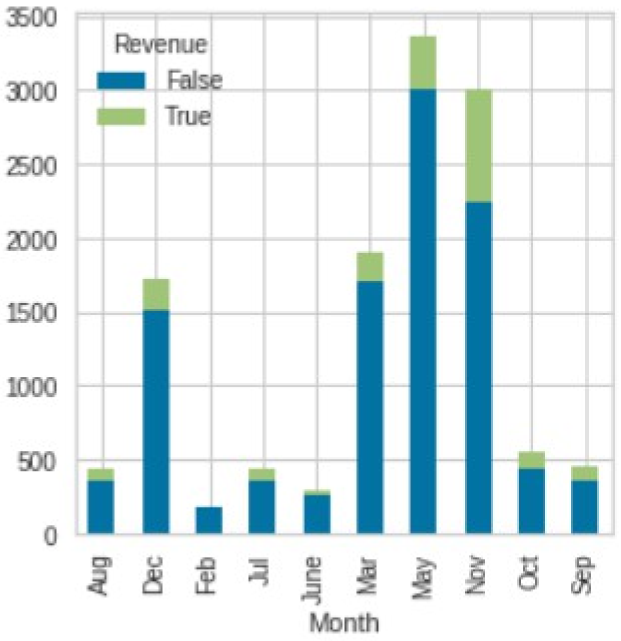
Relation ship of Metrices with Revenue

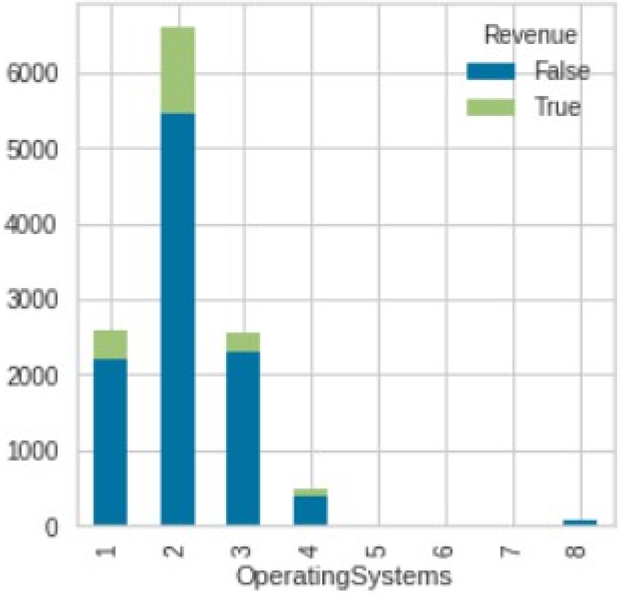


Analysis of Categorical Variable with target variable Revenue :

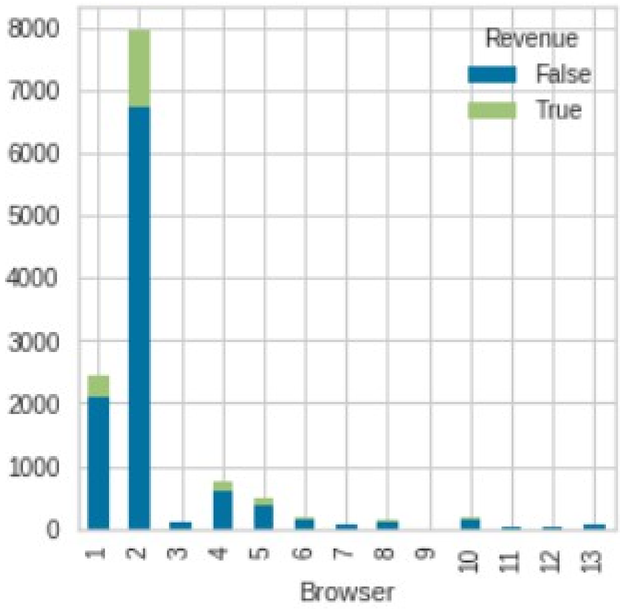


Above graph represent that at special day 0.0 mostly the target is drawn and mostly are False mean purchased ration is low.

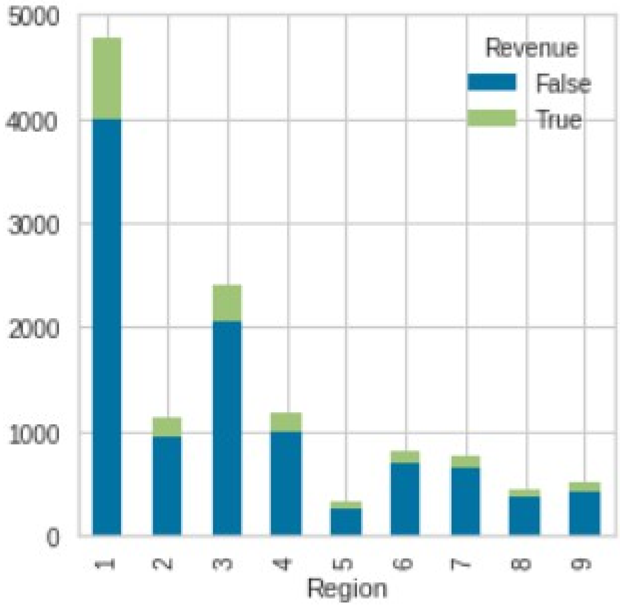




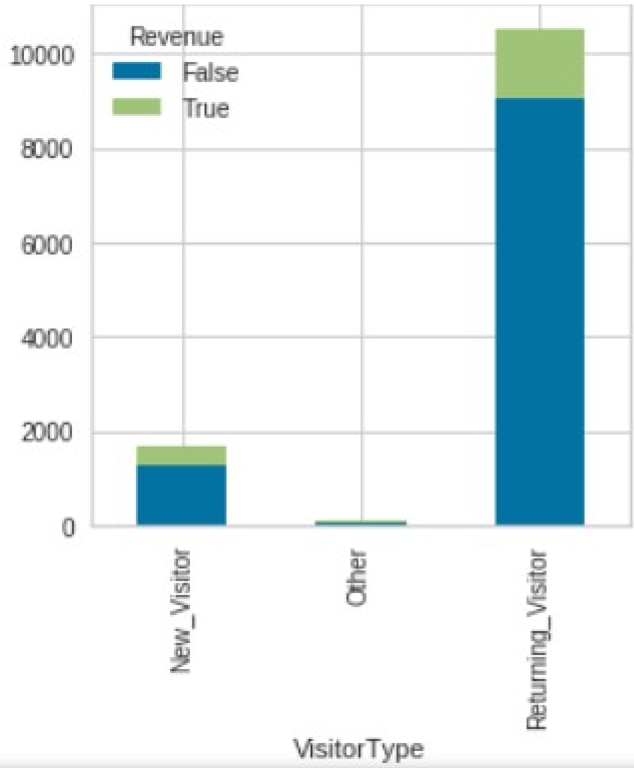
Above graph represent that from operationg system 2,1 and 3 visitor visit the website.and the Revenue is generate greater in amount only from operating system 2.



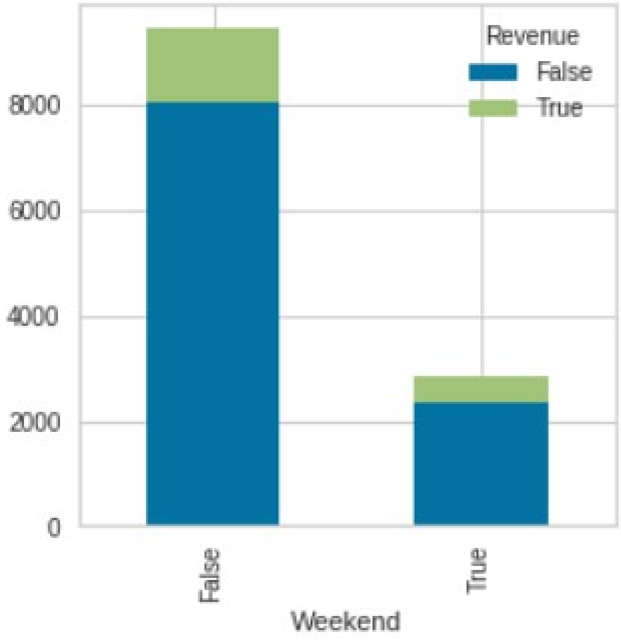
Above graph represent that from browser 2 visitor visit the website and did shopping online but very rare, mostly customer only visit but did not purchased any thing.



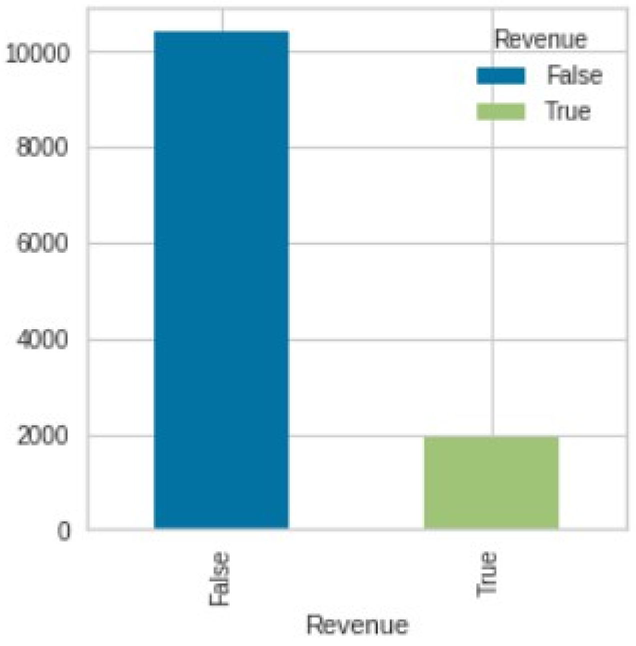
from Region 1 and 3 customer visit the website.



Returning Visitor visit website mostly and did purchased .



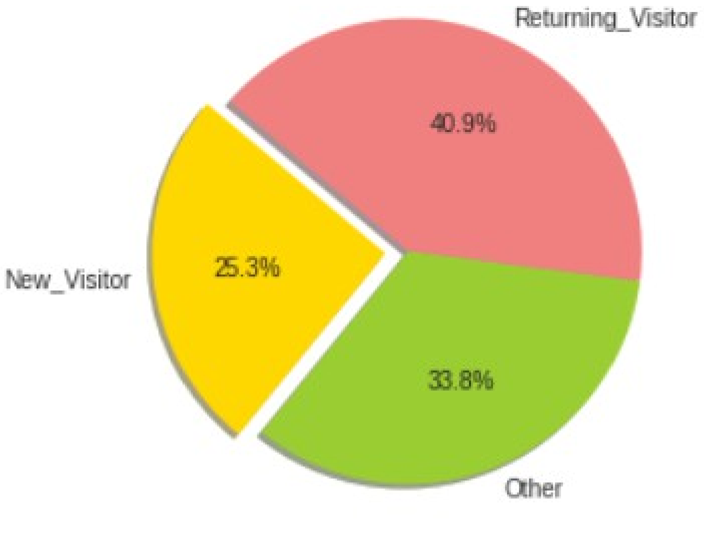
Mostly the website was visited in Odd days.



Above graph Represent that mostly the Revenue is not generated.
The customer only visit the website but did not purchased.

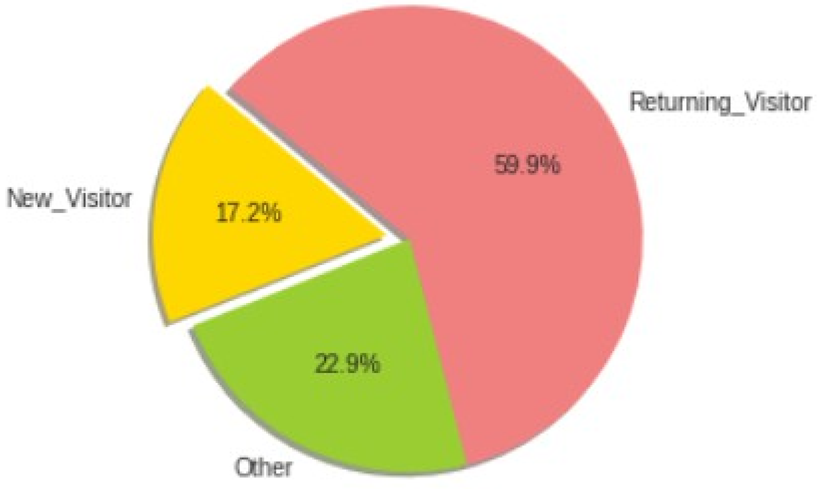
Analyze Visitor Types with Pages Durations :

Relation ship between Visitor type with Administrative_Duration

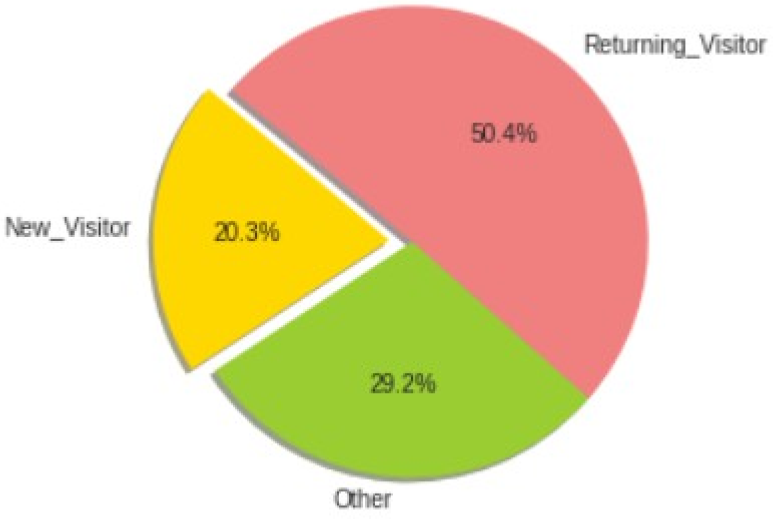


Returning visitor duration time is high in Administrative Duration page

Relation ship between Visitor type with Informational_Duration



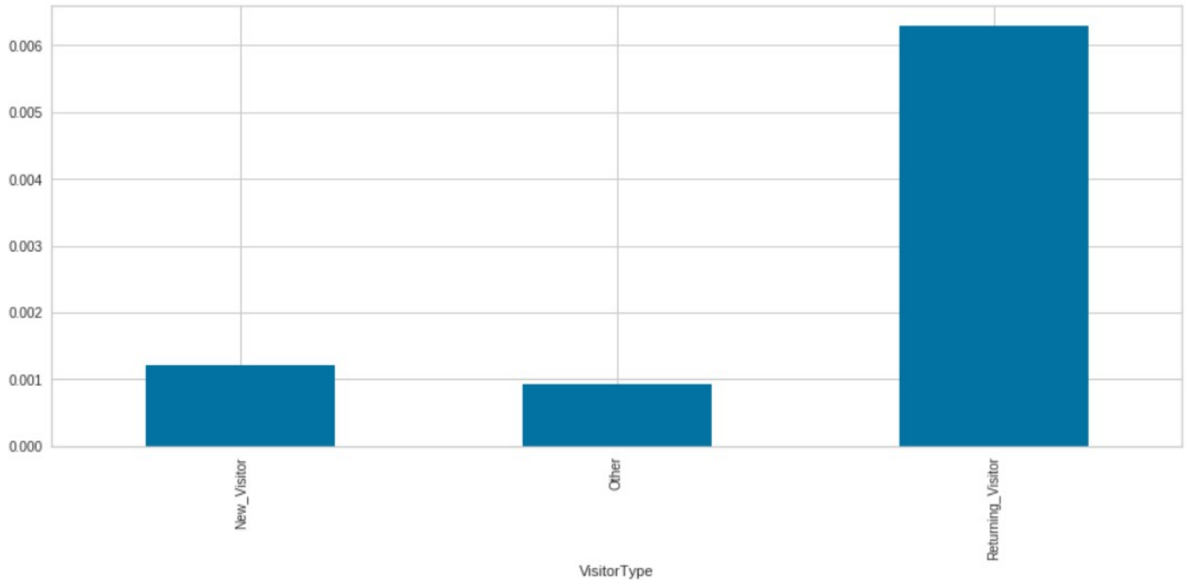
Relation ship between Visitor type with ProductRelated_Duration.



Above relation represent that returning and other durations on webpage are higher.

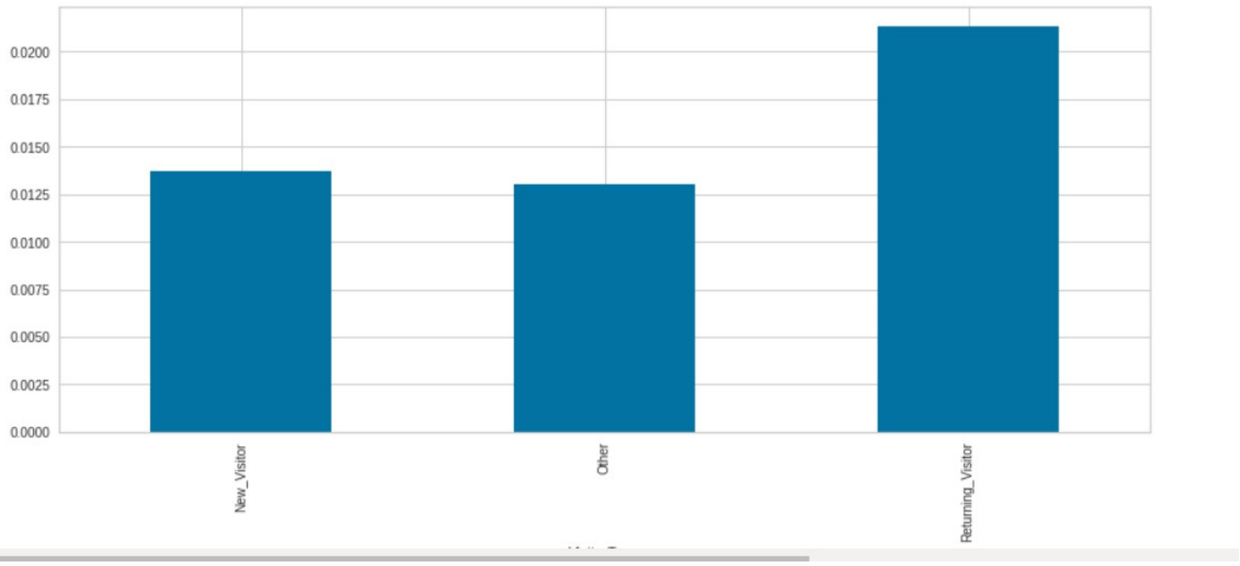
Analyze Visitor Types with Metrics:

With Bounce Rate:



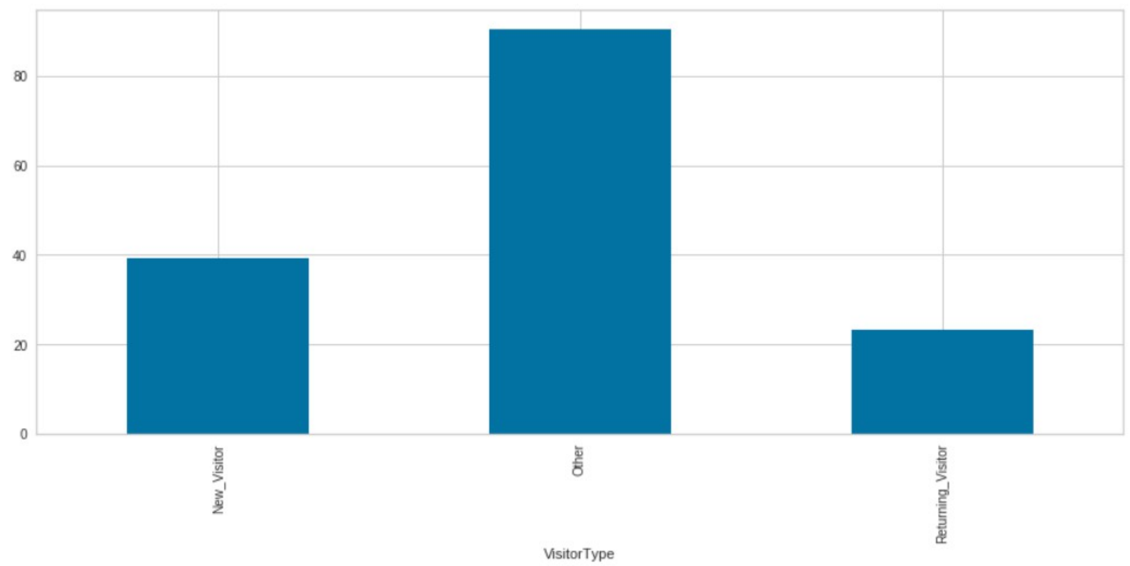
Bounce Rate of Returning Visitor is high.

With Exit Rate:



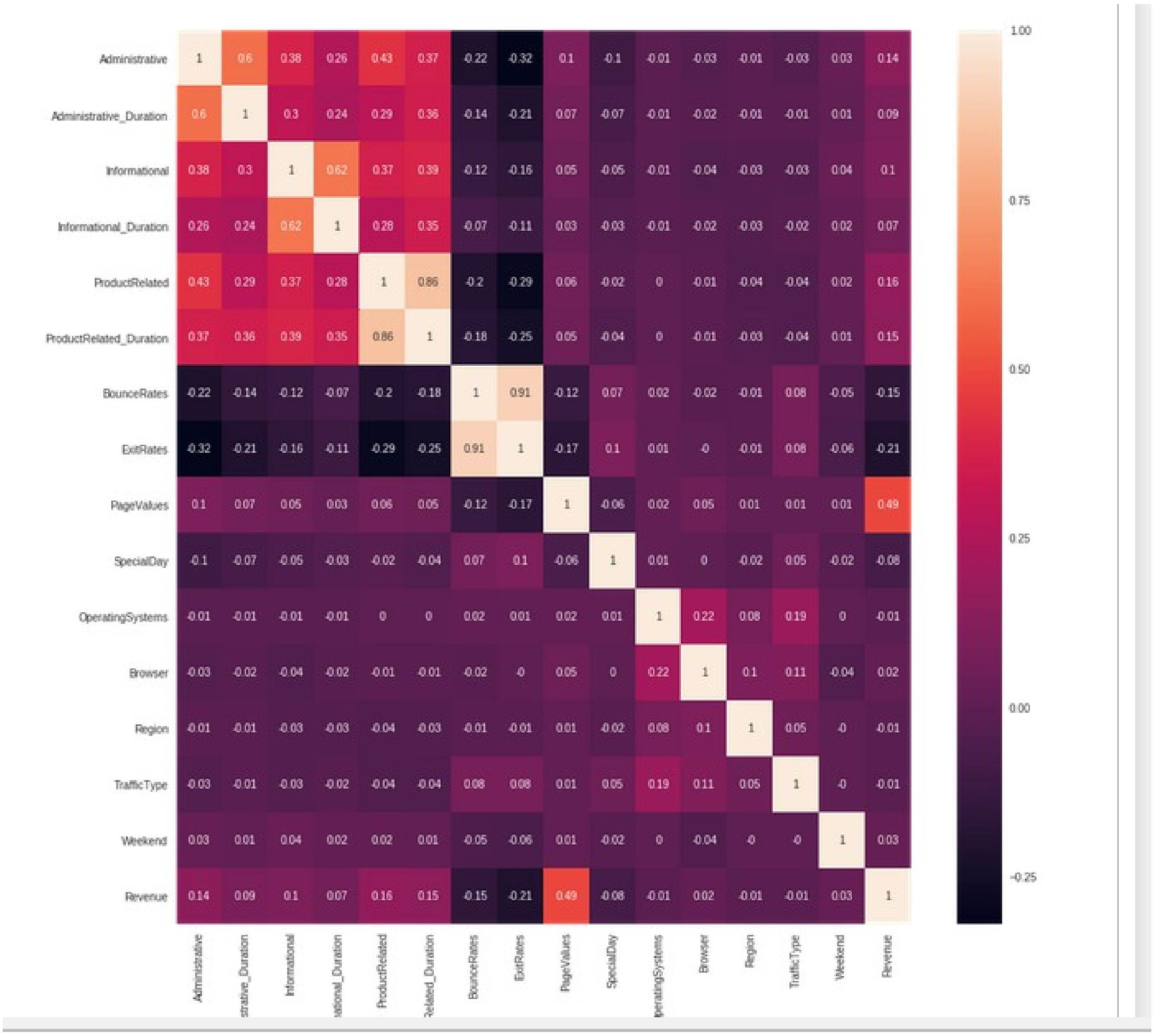
Exit Rate of returning visitor is high and other and new visitor exit rate is same.

With page Value:



The other type of visitor has higher page value.

Correlation :
Find correlated variables and variables affecting the target variable the most.



	Administrative	strative_Duration	Informational	ational_Duration	ProductRelated	Related_Duration	BounceRates	ExitRates	PageValues	SpecialDay	peratingSystems	Browser	Region	TrafficType	Weekend	Revenue
Administrative	1	0.6	0.38	0.26	0.43	0.37	-0.22	-0.32	0.1	-0.1	-0.01	-0.03	-0.01	-0.03	0.03	0.14
Administrative_Duration	0.6	1	0.3	0.24	0.29	0.36	-0.14	-0.21	0.07	-0.07	-0.01	-0.02	-0.01	-0.01	0.01	0.09
Informational	0.38	0.3	1	0.62	0.37	0.39	-0.12	-0.16	0.05	-0.05	-0.01	-0.04	-0.03	-0.03	0.04	0.1
Informational_Duration	0.26	0.24	0.62	1	0.28	0.35	-0.07	-0.11	0.03	-0.03	-0.01	-0.02	-0.03	-0.02	0.02	0.07
ProductRelated	0.43	0.29	0.37	0.28	1	0.86	-0.2	-0.29	0.06	-0.02	0	-0.01	-0.04	-0.04	0.02	0.16
ProductRelated_Duration	0.37	0.36	0.39	0.35	0.86	1	-0.18	-0.25	0.05	-0.04	0	-0.01	-0.03	-0.04	0.01	0.15
BounceRates	-0.22	-0.14	-0.12	-0.07	-0.2	-0.18	1	0.91	-0.12	0.07	0.02	-0.02	-0.01	0.08	-0.05	-0.15
ExitRates	-0.32	-0.21	-0.16	-0.11	-0.29	-0.25	0.91	1	-0.17	0.1	0.01	-0	-0.01	0.08	-0.06	-0.21
PageValues	0.1	0.07	0.05	0.03	0.06	0.05	-0.12	-0.17	1	-0.06	0.02	0.05	0.01	0.01	0.01	0.49
SpecialDay	-0.1	-0.07	-0.05	-0.03	-0.02	-0.04	0.07	0.1	-0.06	1	0.01	0	-0.02	0.05	-0.02	-0.08
OperatingSystems	-0.01	-0.01	-0.01	-0.01	0	0	0.02	0.01	0.02	0.01	1	0.22	0.08	0.19	0	-0.01
Browser	-0.03	-0.02	-0.04	-0.02	-0.01	-0.01	-0.02	-0	0.05	0	0.22	1	0.1	0.11	-0.04	0.02
Region	-0.01	-0.01	-0.03	-0.03	-0.04	-0.03	-0.01	-0.01	0.01	-0.02	0.08	0.1	1	0.05	-0	-0.01
TrafficType	-0.03	-0.01	-0.03	-0.02	-0.04	-0.04	0.08	0.08	0.01	0.05	0.19	0.11	0.05	1	-0	-0.01
Weekend	0.03	0.01	0.04	0.02	0.02	0.01	-0.05	-0.06	0.01	-0.02	0	-0.04	-0	-0	1	0.03
Revenue	0.14	0.09	0.1	0.07	0.16	0.15	-0.15	-0.21	0.49	-0.08	-0.01	0.02	-0.01	-0.01	0.03	1

From the above correlation we notice that

administrative and administrative duration are highly correlated

Informational and Informational Duration are highly correlated

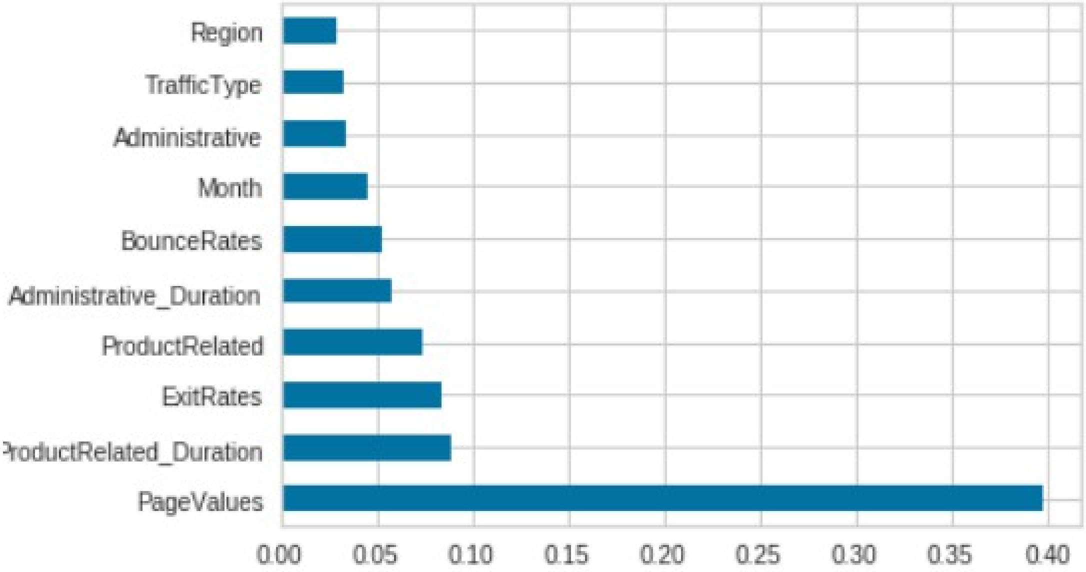
Product and Product duration are highly correlated

Exit and bounce rate are highly correlated

Page Value is highly corelated with target variable Revenue.

Selecting Features for Cluster:

We first check for correlation those variable which are highly correlated with other,we remove them and select one of them..
We label encode the categorical variable,we use random forest for feature selection..from which we get this feature.
1) Page Values.
2) Exit Rate.
3)ProductRelatedDuration.

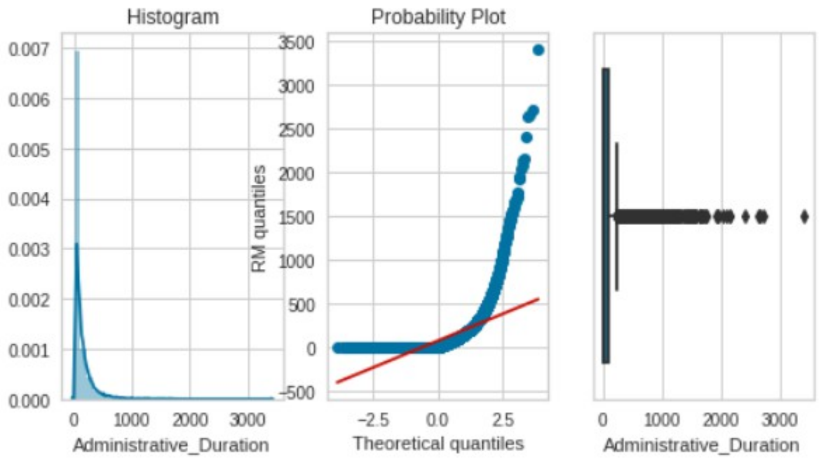


Analyzes For Numerical Variable and Outlier Engineering:

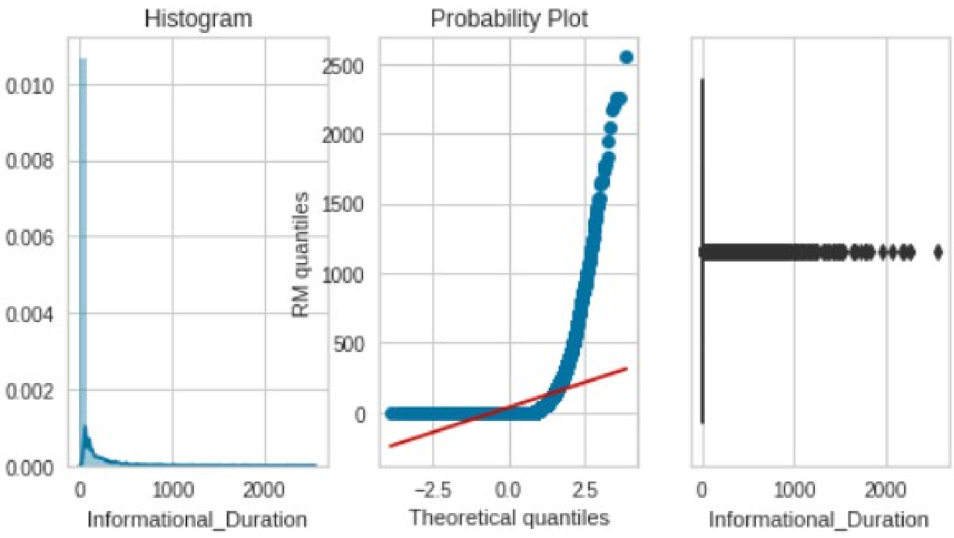
Reasons for using Plots:

We use qq plot to check for linearity of data .
We use Histogram for distplot to check skewness of Features.
We use box plot for outlier Detection.

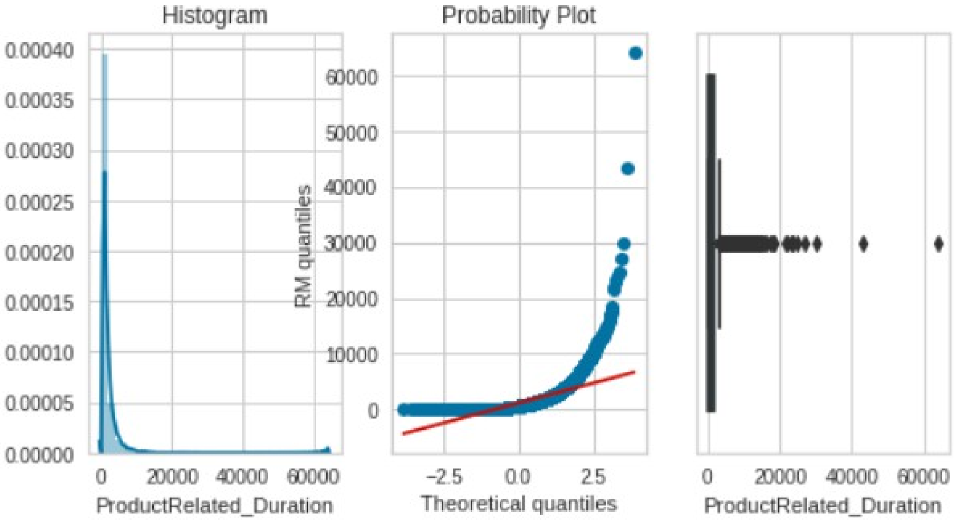
Analyze of Administrative Duration :



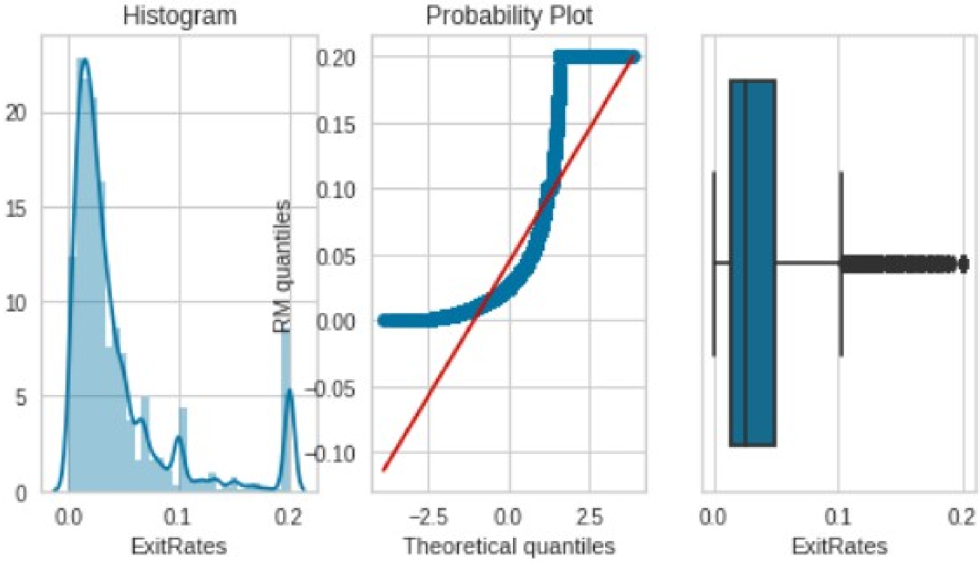
Analyze of Informational Duration Duration :



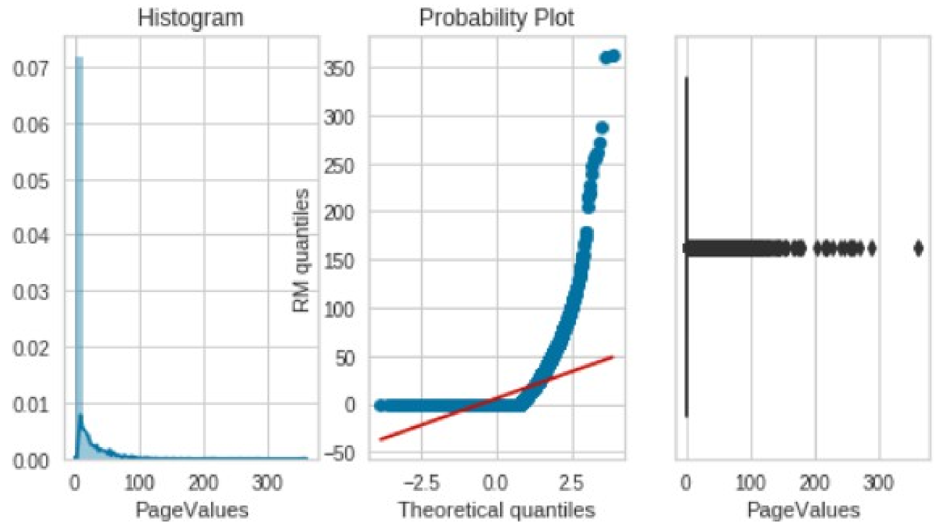
Analyze of Product Related Duration :



Analyze of Exit Rate :

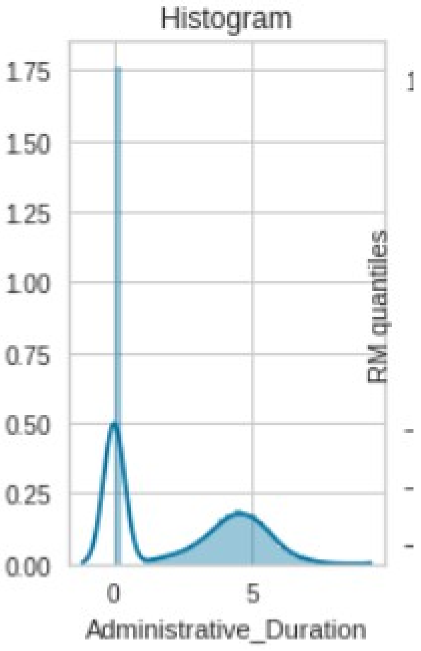


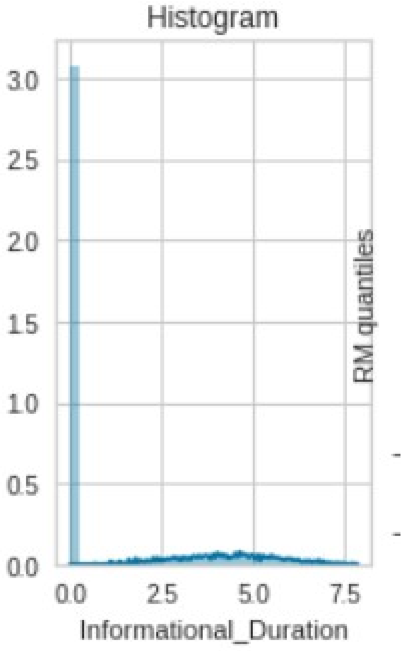
Analyze of Page Value:

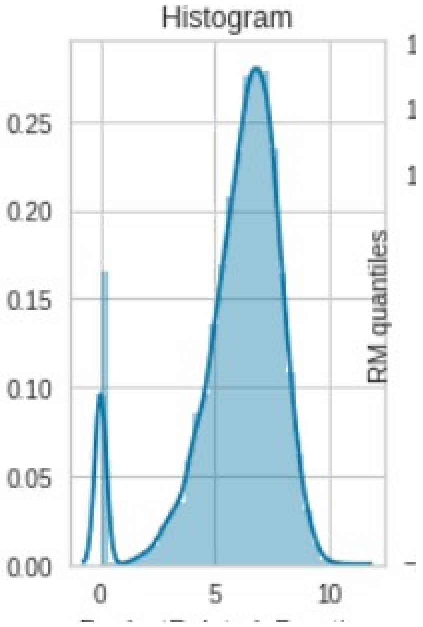


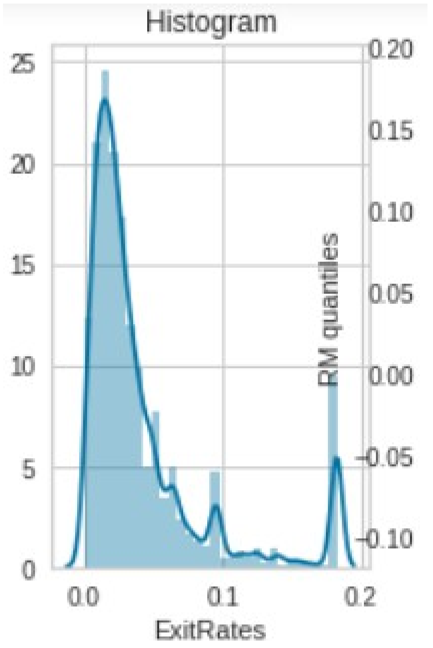
we use log transform to deal to with outliers.

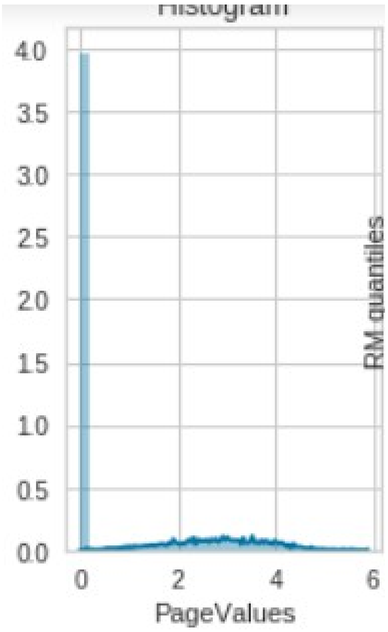
After Transform out Data shape is :





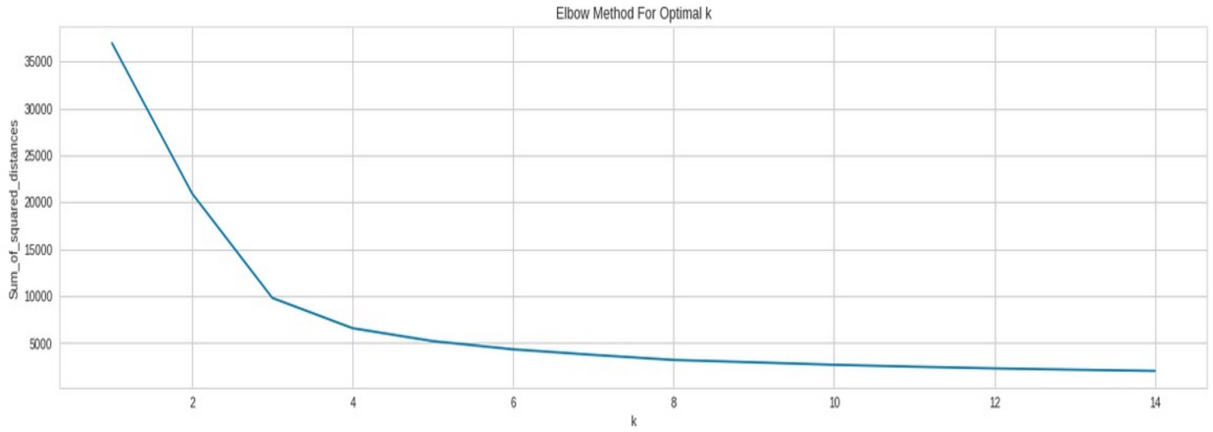






Cluster Graph and Result Table:

We use Elbow Method for K mean for finding for K.



Algorithm	Silhouette Index	Calinski Index	DB Index	External Validation
K means	0.58	17074.41	0.59	0.183
MiniBatch Kmeans with batch size 20	0.55	14153.80	0.64	0.179
MiniBatch Kmeans with batch size 150	0.55	13922.34	0.65	0.175
MiniBatch Kmeans with batch size 200	0.554	14092.94	0.647	0.179
MiniBat	0.55	14205.39	0.64	0.18

ch Kmeans with batch size 250				
MiniBatch Kmeans with batch size 300	0.55	14202.80	0.643	0.180
Mean Shift with bandwidth 1.44	0.58	16725.79	0.567	0.192
Mean Shift with bandwidth 1.17	0.582	16755.54	0.568	0.192
Mean Shift with bandwidth 1.60	0.58	16752.30	0.56	0.19
Mean Shift with bandwidth 1.77	0.58	16792.89	0.56	0.19
Gaussian Mixture Models	0.44	15380.86	0.73	0.129
Gaussian Mixture Models with covariance = Tied	0.501	16297.19	0.68	0.13
Gaussian Mixture	0.47	16047.94	0.705	0.131

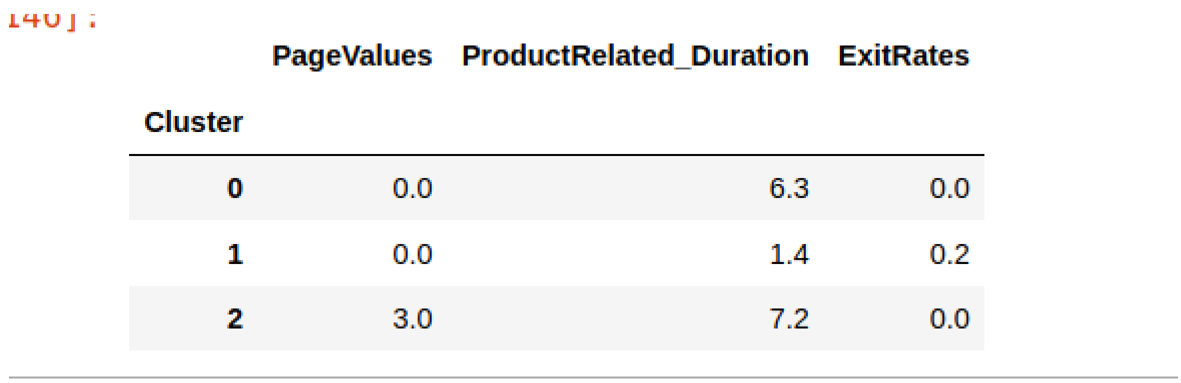
Models with covariance = diag				
Gaussian Mixture Models with spherical	0.50	17920.92	0.66	0.12
DBScan	0.34	16.8	1.13	0.00
DBSCAn using brute Algo	0.34	16.877	1.139	0.000
Optics	0.29	18.68	1.23	0.0139

Result Analysis:

From these models, we can choose the most well segmented model, that is k-means. We use the clusters from the that model to analyze the dataset.

Algorithm K Mean is best among all algorithm it give Silhoutte Index close to 1,Calinski Index is higher,and DB index is lower.

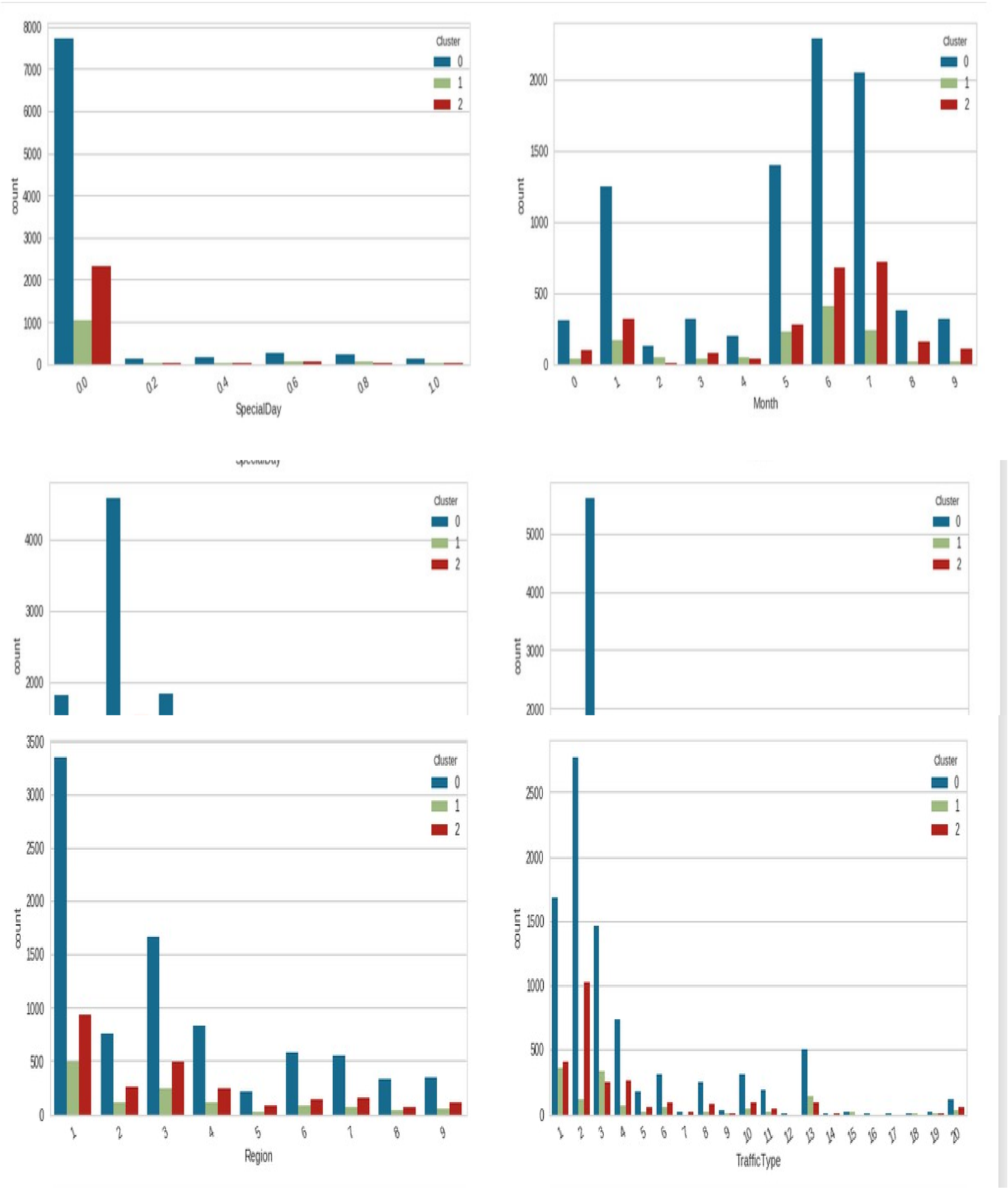
Silhouette Index	0.58
Calinski Index	17074.41
DB Index	0.59
External Validation	0.183

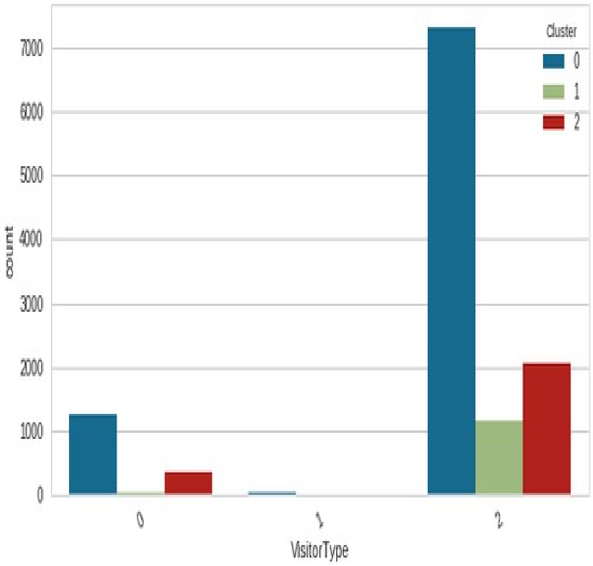


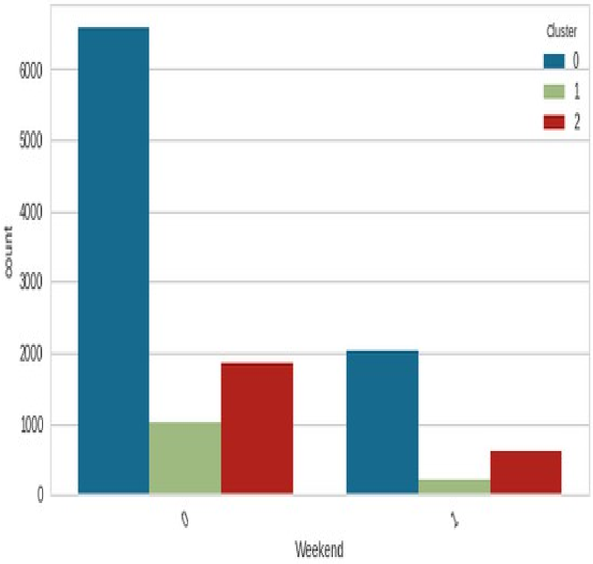
140]:

Cluster	PageValues	ProductRelated_Duration	ExitRates
0	0.0	6.3	0.0
1	0.0	1.4	0.2
2	3.0	7.2	0.0

Assigning cluster with categorical variable.







Summary:

After comparing three kind of clustering models, we decide to use k-means as the model
The data divided into three clusters
The three clusters can be used to determine the visitor purchased or not
Each of the cluster have their own characteristics